Application of Bayesian least absolute shrinkage and selection operator (LASSO) and BayesCπ methods for genomic selection in French Holstein and Montbéliarde breeds

Carine Colombani Colombani, Andres Legarra, Sebastien S. Fritz, François F. Guillaume, Pascal Croiseau, Vincent Ducrocq, Christèle Robert-Granié

J. Dairy Sci. 96:575–591
http://dx.doi.org/10.3168/jds.2011-5225

Application of Bayesian least absolute shrinkage and selection operator (LASSO) and BayesCπ methods for genomic selection in French Holstein and Montbéliarde breeds

C. Colombani,* A. Legarra,* S. Fritz,† F. Guillaume,‡ P. Croiseau,§ V. Ducrocq,§ and C. Robert-Granié*[1]
*INRA, UR631-SAGA, BP 52627, 31326 Castanet-Tolosan Cedex, France
†Union Nationale des Coopératives d'Elevage et d'Insémination Animale (UNCEIA), 149 rue de Bercy, 75595 Paris, France
‡Institut de l'Élevage, 149 rue de Bercy, 75595 Paris, France
§INRA, UMR1313-GABI, 78352 Jouy en Josas, France

ABSTRACT

Recently, the amount of available single nucleotide polymorphism (SNP) marker data has considerably increased in dairy cattle breeds, both for research purposes and for application in commercial breeding and selection programs. Bayesian methods are currently used in the genomic evaluation of dairy cattle to handle very large sets of explanatory variables with a limited number of observations. In this study, we applied 2 Bayesian methods, BayesCπ and Bayesian least absolute shrinkage and selection operator (LASSO), to 2 genotyped and phenotyped reference populations consisting of 3,940 Holstein bulls and 1,172 Montbéliarde bulls with approximately 40,000 polymorphic SNP. We compared the accuracy of the Bayesian methods for the prediction of 3 traits (milk yield, fat content, and conception rate) with pedigree-based BLUP, genomic BLUP, partial least squares (PLS) regression, and sparse PLS regression, a variable selection PLS variant. The results showed that the correlations between observed and predicted phenotypes were similar in BayesCπ (including or not pedigree information) and Bayesian LASSO for most of the traits and whatever the breed. In the Holstein breed, Bayesian methods led to higher correlations than other approaches for fat content and were similar to genomic BLUP for milk yield and to genomic BLUP and PLS regression for the conception rate. In the Montbéliarde breed, no method dominated the others, except BayesCπ for fat content. The better performances of the Bayesian methods for fat content in Holstein and Montbéliarde breeds are probably due to the effect of the *DGAT1* gene. The SNP identified by the BayesCπ, Bayesian LASSO, and sparse PLS regression methods, based on their effect on the different traits of interest, were located at almost the same position on the genome. As the Bayesian methods resulted in regressions of direct genomic values on daughter trait deviations closer to 1 than for the other methods tested in this study, Bayesian methods are suggested for genomic evaluations of French dairy cattle.
Key words: genomic selection, Bayesian method, variable selection, Holstein and Montbéliarde breeds

INTRODUCTION

In recent years, massive amounts of SNP marker data have been made available in dairy cattle for application in selection schemes. In the future, the increase of the density of SNP data will be ensured by the rapid decrease in genotyping costs. However, the number of genotyped and phenotyped animals that constitute reference populations remains limited. Reference populations provide the prediction equations that give genomic EBV (**GEBV**). Genomic EBV are obtained through the estimation of SNP effects in a context where the number of independent variables (SNP markers) is much larger than the number of individuals of the reference population.

In the literature, several methods have been proposed to estimate SNP effects assuming or not a prior distribution of SNP effects (Bayesian vs. frequentist methods). The Bayesian methods differentiate in the assumed prior distributions of SNP effects.

For the estimation of SNP effects, Meuwissen et al. (2001) proposed 2 Bayesian methods, named BayesA and BayesB. The BayesA method assumes that the prior distribution of the SNP effects is a normal distribution with a 0 mean and a different variance for each SNP. The prior distribution of these variances in BayesA is proportional to a scaled inverted chi-squared distribution, noted $\chi^{-2}(\nu,S)$, where ν = degrees of freedom and S = a scale parameter. In the BayesB method, a stochastic search variable selection is used, which assumes that only part of the SNP involved provide information about the phenotype. A combination of normal

Received December 5, 2011.
Accepted September 14, 2012.
[1]Corresponding author: Christele.Robert-Granie@toulouse.inra.fr

distribution with a 0 mean and a large variance (with probability π) and a distribution with point mass only at zero (with probability $1 - \pi$) is assigned to each SNP effect. Both BayesA and BayesB assume a Student's *t*-distribution at the level of SNP effects (Sorensen and Gianola, 2002). Since then, BayesA and BayesB methods have been widely used in animal breeding research. Several studies have also used a simple BLUP approach [also often referred to as genomic BLUP (**GBLUP**) or SNP-BLUP], as described in Meuwissen et al. (2001) as a reference method, to compare the gain in accuracy with Bayesian methods. BayesA and BayesB were shown to be similar or slightly more reliable than GBLUP in Australian Holstein-Friesian bulls (Hayes et al., 2009) and in the New Zealand reference population (Harris et al., 2009). Using a Fleckvieh reference population (Gredler et al., 2009), BayesB was found to be more accurate for 3 traits out of 4 than the BayesA method modified to include a polygenic effect (Hayes, 2009).

The BayesC model (Kizilkaya et al., 2010) differs from BayesB by using a common variance for SNP with a nonzero effect, instead of a locus-specific variance. This variance is estimated, in contrast to GBLUP, where it is supposed as known. Habier et al. (2011) extended the panel of Bayesian methods with BayesCπ, treating the probability π that a SNP marker has an effect as an unknown parameter, which can be estimated. BayesCπ was compared with BayesA and B using simulated and real data from North American Holstein bulls. The results showed that the accuracies of GEBV were similar for the different methods.

The Bayesian least absolute shrinkage and selection operator (**LASSO**) method was also used in a genomic evaluation context (de los Campos et al., 2009; Weigel et al., 2009), but these studies did not compare Bayesian LASSO with other genomic selection methods. de los Campos et al. (2009) compared the predictive ability of different Bayesian LASSO models with respect to the choice of prior for the regularization parameters on simulated data, using pedigree information only, marker information only, or considering pedigree and marker information jointly, in wheat line data sets and populations of mice. The results showed that a double-exponential prior may be a better choice than a Student's *t*-distribution prior (such as BayesA) if most markers do not have any effect. They outlined that a Student's *t*-distribution may place more density at zero than the Gaussian prior of standard Bayesian methods (the density at zero is larger in the double-exponential prior). The Bayesian LASSO appears to be an interesting alternative to the BayesA method for performing regressions on markers. They also have shown, on real data sets, that the model with both a polygenic and SNP effect was the most efficient. Legarra et al. (2011) and Ostersen et al. (2011) showed that Bayesian LASSO and GBLUP gave comparable results for most traits, on real data sets of Montbéliarde and Holstein bulls, and on Danish Duroc pigs, respectively.

Gredler et al. (2009) used partial least squares (**PLS**) regression on a Fleckvieh reference population and they compared it with BayesA, BayesB, and GBLUP. The PLS method reduces the dimension of the regression model by building orthogonal linear combinations of markers or components that have a maximal correlation with the trait. The PLS regression and GBLUP gave similar results but with lower accuracies than those obtained with BayesB and higher accuracies than with BayesA. Genomic BLUP was also shown to be similar to PLS regression for 2 traits with dairy bull data (Moser et al., 2009) and for 3 traits with French Lacaune dairy sheep data (Robert-Granié et al., 2011) and very slightly better than PLS regression in the Fleckvieh breed (Gredler et al., 2009). Colombani et al. (2010) have shown that PLS regression and sparse PLS (**sPLS**) regression (method performing variable selection in addition to reducing dimensionality) provided the same correlations as GLUP for 4 traits with French Holstein bulls.

The main goal of this study was to compare BayesCπ and Bayesian LASSO with methods currently used in dairy cattle evaluation, such as pedigree-based BLUP and GBLUP, and methods recently used in genomic selection by Long et al. (2011) and Colombani et al. (2012) and known to perform well with large data sets (Lê Cao et al., 2008; Chun and Keles, 2009) such as PLS regression and sPLS regression. After studying the statistical modeling and the convergence properties of BayesCπ, we compared the different methods, based on their predictive abilities, using 2 real data sets from Montbéliarde and Holstein breeds. Then, the positions of SNP selected by BayesCπ, Bayesian LASSO, and sPLS regression were compared.

MATERIALS AND METHODS

Data

Data sets consisted of 1,172 Montbéliarde bulls and 3,940 French Holstein bulls, progeny tested and genotyped with the Illumina Bovine SNP50K BeadChip (Illumina Inc., San Diego, CA). Training and validation data sets were defined for each breed, according to a cutoff birth date defined so that the validation set included the youngest 25% genotyped bulls. Consequently, 2 training data sets consisting of 950 Montbéliarde bulls and 2,976 Holstein bulls were available to

provide prediction equations for both breeds and the 3 traits (milk yield, fat content, and conception rate). Next, the phenotypes of the 222 Montbéliarde bulls and 964 Holstein bulls from the validation data sets, born between June 2002 and 2004, were predicted. Pedigree files for the 2 breeds included 4,717 and 12,142 bulls in the Montbéliarde and Holstein breeds, respectively.

The DualPHASE software (Druet and Georges, 2010) was used to check Mendelian segregation and infer missing genotypes from large-family information. Minimum minor allele frequencies of 3% were required and resulted in 38,462 SNP for the Montbéliarde breed and 39,738 SNP for the Holstein breed, used as independent variables.

The response variables (phenotypes) were daughter yield deviations (**DYD**; VanRaden and Wiggans, 1991; Mrode and Swanson, 2004) from the October 2009 national evaluation. The precision of DYD was accounted for through the weighting of DYD by their error variance, which is a function of the sire's effective daughter contribution (**EDC**). Three traits were considered with different heritabilities (h^2): milk yield ($h^2 = 0.3$), fat content ($h^2 = 0.5$), and conception rate ($h^2 = 0.02$; Boichard and Manfredi, 1994).

Genomic BLUP

Two methods were used as reference methods to assess the predictive ability of PLS regression, sPLS regression, Bayesian LASSO, and BayesCπ: pedigree-based BLUP and GBLUP. The general statistical model was $\mathbf{y} = \mu\mathbf{1} + \mathbf{Za} + \mathbf{e}$, where $\mathbf{y}$ is a vector of phenotypes (DYD), μ is the overall mean, $\mathbf{Z}$ is a design matrix allocating observations to breeding values, $\mathbf{a}$ is a random vector of additive genetic values, and $\mathbf{e}$ is a vector of random normal errors. In pedigree-based BLUP, $\mathrm{Var}(\mathbf{a}) = \mathbf{A}\sigma_a^2$, where $\mathbf{A}$ is the pedigree-based relationship matrix and σ_a^2 is the additive genetic variance. In GBLUP, genomic information was included in the BLUP model assuming a prior normal distribution for SNP markers (VanRaden, 2008) and using mixed-model equations with a genomic relationship matrix (Cole et al., 2009; VanRaden et al., 2009). Using genomic information implies that the relationship matrix $\mathbf{A}$ based on pedigree is substituted by the genomic relationship matrix ($\mathbf{G}$) as defined by VanRaden (2008). We assumed that $\mathrm{Var}(\mathbf{a}) = \mathbf{G}\sigma_a^2$, with

$$\mathbf{G} = \frac{\mathbf{XX}'}{2\sum_{j=1}^{p} q_j\left(1-q_j\right)},$$

where p is the number of loci considered, q_j is the frequency of an allele of the marker j, and $\mathbf{X}$ is a centered incidence matrix of SNP effects, corrected for allele frequencies.

PLS and sPLS Regressions

PLS Regression. The PLS regression (Wold, 1966) is a dimension reduction method developed to deal with the "$p \gg n$" problem (the number of predictors p is much larger than the number of observations n). It combines principal components analysis and multiple regressions to handle very large sets of independent variables, which can be highly correlated, such as our set of SNP predictors.

Partial least squares regression relies on successive regressions of the response variable $\mathbf{y}$ onto substitutes of the initial independent variables ($\mathbf{X}$), named latent variables, which define a space of smaller dimension. The latent variables ($\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_H$) are linear combinations of the independent variables ($\mathbf{X}$) through loadings vectors ($\mathbf{u}_1, \ldots, \mathbf{u}_H$), where H is the number of latent variables retained in the final PLS regression model. These parameters are estimated to solve the following optimization problem:

$$\max_{\|\mathbf{u}_h\|=1} \mathrm{cov}\left(\mathbf{X}_{h-1}\mathbf{u}_h, \mathbf{y}_{h-1}\right),$$

where $\boldsymbol{\xi} = \mathbf{X}_{h-1}\mathbf{u}_h$ and $\mathbf{X}_h$ and $\mathbf{y}_h$ are the residual matrices of the regression of $\mathbf{X}_{h-1}$ and $\mathbf{y}_{h-1}$ onto $\boldsymbol{\xi}_h$ for each PLS regression dimension $h = 1, \ldots, H$, where $\mathbf{X}_0 = \mathbf{X}$ and $\mathbf{y}_0 = \mathbf{y}$.

The parameter H can be tuned by cross-validation, as proposed by Chun and Keleş (2009) and Coster et al. (2010). Solberg et al. (2009) proposed an alternative to fix the parameter H to obtain the PLS regression prediction equation which leads to the highest correlation between observed and predicted phenotypes from the validation data set. Colombani et al. (2012) tested and discussed these 2 approaches. Following their approach, 39, 24, and 7 latent variables provided optimal models for milk yield, fat content, and conception rate, respectively, in the Montbéliarde data set. In the Holstein data set, 42, 83, and 29 latent variables were included in the best PLS regression models for milk yield, fat content, and conception rate, respectively.

sPLS Regression. Sparse PLS regression, developed by Lê Cao et al. (2008) and later by Chun and Keleş (2009), differs from PLS regression by adding a step of variable selection to each latent variable through the loading vectors ($\mathbf{u}_1, \ldots, \mathbf{u}_H$). The sparsity of the loading vectors is introduced iteratively by penalizing

$\mathbf{u}_h$ with a soft-thresholding penalization, as in sparse principal components analysis (Shen and Huang, 2008). The optimization problem becomes

$$\max_{\|\mathbf{u}_h\|=1} \operatorname{cov}\left(\mathbf{X}_{h-1}\mathbf{u}_h, \mathbf{y}_{h-1}\right) + g_\lambda\left(\mathbf{u}_h\right),$$

where $g_\lambda(\mathbf{x}) = sign(\mathbf{x})(|\mathbf{x}| - \lambda)_+$ is the soft-thresholding penalty function of the $\mathbf{x}$ vector, and λ represents the intensity of penalization.

The number of dimensions (H) is fixed as in PLS regression: 24, 20, and 2 latent variables in the Montbéliarde breed and 44, 50, and 27 latent variables in the Holstein breed for milk yield, fat content, and conception rate, respectively.

The sparsity is set through the choice of the number of selected variables in each dimension. This choice can be made by examining the root mean squared error of prediction (**RMSEP**) with K-fold cross-validation ($K = 10$) within the training data set and for each given dimension h (Mevik and Cederkvist, 2004):

$$\text{RMSEP} = \sqrt{\frac{1}{K}\sum_{k=1}^{K}(\hat{\mathbf{y}}_k - \mathbf{y}_k)},$$

where $\mathbf{y}_k$ and $\hat{\mathbf{y}}_k$ are the vectors of observed and predicted DYD, respectively. The adequate number of selected variables is the one that minimizes RMSEP. In the Montbéliarde breed and for milk yield, 10% of the initial number of SNP were kept in each of the 24 dimensions, which amounts to 28,837 SNP in the final model. For the other traits, the corresponding figures were 3% for a total number of 14,447 SNP for fat content and 5% for 3,808 SNP for conception rate. In the Holstein breed, 4, 0.8, and 4% of the initial number of SNP (that is, 22,948 SNP, 9,832 SNP, and 20,150 SNP) were kept for milk yield, fat content, and conception rate, respectively (Colombani et al., 2012). Partial least squares and sPLS regressions were performed with the R package mixOmics (previously named integrOmics; R Foundation for Statistical Computing, Vienna, Austria; Lê Cao et al., 2009).

Evaluation of SNP Effects in PLS and sPLS Regressions. After fitting the PLS and sPLS regression models, we obtained a vector of regression coefficients with respect to the original variables, which could be directly used for prediction. As a result of variable selection in the sPLS regression model, some of the estimated coefficients were exactly zero. To assess marker contributions, variable importance in projection (***VIP***) coefficients were used to measure the contribution of each SNP x_j in the construction of $\mathbf{y}$ through latent variables $\boldsymbol{\xi}_h$. A *VIP* coefficient was defined for each SNP x_j and for a model with H dimensions by

$$VIP_{Hj} = \sqrt{\frac{p}{\sum_{h=1}^{H} cor^2(\mathbf{y}, \boldsymbol{\xi}_h)} \sum_{h=1}^{H} cor^2(\mathbf{y}, \boldsymbol{\xi}_h) w_{hj}^2},$$

with

$$\sum_{j=1}^{p} VIP_{Hj}^2 = p.$$

The contribution of x_j in the construction of $\boldsymbol{\xi}_h$ was measured by its weight w_{hj}, provided by PLS or sPLS regression. Because the mean of squared *VIP* scores equals 1, the greater-than-1 rule is generally used as a criterion for variable selection.

Bayesian LASSO

In a genomic selection context, Legarra et al. (2011) proposed a general model for Bayesian LASSO equivalent to the original LASSO proposed by Tibshirani (1996) by splitting the sources of variation in a purely residual term $\left(\sigma_e^2\right)$ and variation due to SNP $\left(\sigma_g^2\right)$. It can be seen as a hierarchical model, in which individual variances for each SNP effect are modeled upon a common exponential distribution. In this study, we applied Bayesian LASSO as defined by Legarra et al. (2011), where it was called BL2Var because it was the most accurate method for prediction and accommodated well major genes in their study, on similar data. The model considered is

$$\mathbf{y} = \mu\mathbf{1} + \mathbf{X}\mathbf{g} + \mathbf{e},$$
$$\text{with } \mathbf{g} \mid \lambda \sim \prod_j \frac{\lambda}{2}\exp\left(-\lambda\left|g_j\right|\right) \text{ and } \mathbf{e} \mid \sigma_e^2 \sim MVN\left(\mathbf{0}, \mathbf{I}\sigma_e^2\right),$$

where $\mathbf{y}$ is the vector of phenotypes of the n individuals of the training data set, μ is the overall mean, $\mathbf{X}$ is a ($n \times p$) design matrix that consists of the genotypes of p SNP markers for each of the n individuals, $\mathbf{g} = \{g_j\}$ is the random vector of SNP effect, $\mathbf{e}$ is a random vector of residual effects, and λ is the "sharpness" parameter. The parameterization of SNP genotypes (elements of $\mathbf{X}$) is as in VanRaden (2008): $-2q_j$, $1 - 2q_j$, and $2 - 2q_j$ for the genotypes 00, 01, and 11, respectively, where q_j is the allelic frequency of 1. The prior distribution for the residual variance was an inverted chi-squared distribution with 4 degrees of freedom and expectations

equal to the value used in the regular genetic evaluation for σ_e^2. The prior for λ was considered vague, being uniform between 0 and 1,000,000.

BayesCπ

BayesCπ Model. The last method tested in this study was BayesCπ. It derives from the BayesC method (Kizilkaya et al., 2010; Sun et al., 2011). The statistical model was again

$$\mathbf{y} = \mu\mathbf{1} + \mathbf{Xg} + \mathbf{e},$$

as in the Bayesian LASSO method.

BayesC modifies the BayesB method by replacing the locus-specific variance components by a common effect variance. BayesCπ is equivalent to the BayesC model with an unknown fraction π [with uniform (0, 1) prior] of SNP with a nonzero effect. The probability π is defined so that the prior distribution for the additive SNP effect is zero, with a probability π and normal with a probability $(1 - \pi)$ so that $g_j \left| \pi, \sigma_g^2 = 0 \right.$, with probability π and $g_j \left| \pi, \sigma_g^2 \sim N\left(0, \sigma_g^2\right) \right.$, with probability $(1 - \pi)$.

The variance σ_g^2 was assumed to have a scaled inverse chi-squared prior with ν_g degrees of freedom and scale S_g^2. The marginal prior of $g_j \left| \nu_g, S_g^2 \right.$ was a univariate Student's t-distribution $t\left(0, \nu_g, S_g^2\right)$, with a probability $(1 - \pi)$. As suggested by Habier et al. (2011), we took $\nu_g = 4.2$; S_g^2 was equal to $E[\sigma_g^2](\nu_g - 2)/\nu_g$, where $E[\sigma_g^2] = \tilde{\sigma}_g^2$, and $\tilde{\sigma}_g^2$ was the variance of the additive effect for a randomly sampled locus. As defined in the previous LASSO model, the prior distribution for the residual variance was an inverted chi-squared distribution.

BayesCπPED Model. The BayesCπPED model differs from the previous one (BayesCπ model) by the addition of a polygenic effect, as proposed by Habier et al. (2011). The statistical model became

$$\mathbf{y} = \mu\mathbf{1} + \mathbf{Zu} + \mathbf{Xg} + \mathbf{e},$$

with the same definitions as previously given and where $\mathbf{u}$ is a random vector of the polygenic effects of all the individuals in the pedigree and $\mathbf{Z}$ is an incidence matrix of the polygenic effects. The prior of $\mathbf{u} \left| \mathbf{A}, \sigma_u^2 \right.$ was normal, with mean 0 and variance-covariance matrix $\mathbf{A}\sigma_u^2$, with $\mathbf{A}$ the numerator relationship matrix and σ_u^2 the additive genetic variance not explained by SNP. The prior distribution for σ_u^2 was an inverted chi-squared distribution, as defined by Habier et al. (2011).

Estimation of Variance Components in Bayesian LASSO and BayesCπ

Markov chain Monte Carlo (**MCMC**) was used to estimate the posterior distribution of variances and the model parameters μ, $\mathbf{u}$, σ_u^2 (in the BayesCπPED model), g_j, and σ_g^2, σ_e^2, and π (if unknown). A burn-in period of 20,000 cycles was run before saving results every 50 cycles out of 180,000. The starting value for π was 0.5.

The genetic variance in the population σ_u^2, estimated using a pedigree-based BLUP model, is proportional to the variance of SNP effects σ_g^2 (Gianola et al., 2009). For the LASSO model, the relation is

$$\sigma_g^2 = \frac{\sigma_u^2}{2\sum_{j=1}^{p} q_j(1 - q_j)},$$

and for the 2 Bayes$C\pi$ models,

$$\sigma_g^2 = \frac{\sigma_u^2}{(1-\pi)2\sum_{j=1}^{p} q_j(1 - q_j)},$$

where p is the number of loci considered and q_j is the frequency of an allele of the marker j. Bayesian LASSO and the 2 BayesCπ methods were performed using the GS3 software developed by Legarra et al. (2011; http://snp.toulouse.inra.fr/~alegarra).

Comparison of Methods

The predictive ability of the different methods was compared by considering the EDC-weighted correlation between the observed DYD and predicted DYD from the validation data sets, and the EDC-weighted regression slopes of observed DYD onto predicted DYD from the validation data sets. Ideally, values near 1 were expected (Meuwissen et al., 2001), indicating that the GEBV are unbiased. Marker contributions for each method were also measured via genetic standard deviation units for Bayesian LASSO and BayesCπ and *VIP* coefficients for PLS and sPLS regressions.

The Hotelling-Williams procedure was used to test the difference between the correlations of the different methods. It tests the null hypothesis of equality between 2 dependent correlations that share a variable (Steiger, 1980; Van Sickle, 2003). Under the null hypothesis, the statistical test is distributed as t with n − 3 degrees of freedom. All the correlations discussed

Table 1. Correlations between observed daughter yield deviations (DYD) and predicted DYD in the validation data set without (BayesCπ model) or with (BayesCπPED model) adding a polygenic component to the BayesCπ model

	Montbéliarde[1]		Holstein[2]	
Item	BayesCπ model[3]	BayesCπPED model[4]	BayesCπ model[3]	BayesCπPED model[4]
Milk yield	0.44	0.44	0.57	0.57
Fat %	0.63	0.62	0.80	0.78
Conception rate	0.42	0.44	0.34	0.34

[1]Montbéliarde: training set = 950 bulls; validation set = 222 bulls.

[2]Holstein: training set = 2,976 bulls; validation set = 964 bulls.

[3]BayesCπ model: $\mathbf{y} = \mu\mathbf{1} + \mathbf{Xg} + \mathbf{e}$, where $\mathbf{y}$ is the vector of phenotypes, μ is the overall mean, $\mathbf{X}$ is a design matrix, $\mathbf{g}$ is the random vector of SNP effect, and $\mathbf{e}$ is a random vector of residual effects.

[4]Bayes CπPED model: $\mathbf{y} = \mu\mathbf{1} + \mathbf{Zu} + \mathbf{Xg} + \mathbf{e}$, where $\mathbf{Z}$ is an incidence matrix of the polygenic effects and $\mathbf{u}$ is a random vector of the polygenic effects of all the individuals in the pedigree.

in this study were compared with one another using the Hotelling-Williams test with a 5% threshold.

RESULTS

Considering a Polygenic Effect in the BayesCπ Model

Table 1 shows the correlations between observed DYD and predicted DYD in the validation data set, for both breeds and for the 3 studied traits. It compares the accuracy of a BayesCπ model with only marker effects (BayesCπ model) and a BayesCπ model with marker and polygenic effects (BayesCπPED model). The correlations were not significantly different between the 2 models, considering the Hotelling-Williams test with a threshold of 5% for both breeds and for all of the traits.

Figure 1 presents the regression slopes (b) of observed DYD onto predicted DYD from the validation data set for the same BayesCπ models (BayesCπ model in black and BayesCπPED model in white), the 2 breeds (Montbéliarde on the left and Holstein on the right) and the 3 traits studied (represented on the x-axis). A value of 1 was expected and is depicted with a horizontal line. The confidence intervals were represented by a vertical line at each point and were calculated by adding or subtracting 2 times the standard error. The standard errors were similar for the 2 models and equal to 0.03, 0.02, and 0.07 for milk yield, fat content, and conception rate, respectively, in the Holstein breed, and they were stronger in the Montbéliarde breed (0.10, 0.07, and 0.20 for milk yield, fat content, and conception rate, respectively). The confidence intervals should contain 1. This was the case only in the Montbéliarde breed, for fat content BayesCπPED model ($b = 0.89$), and conception rate BayesCπ model ($b = 1.31$).

The differences between the slopes of BayesCπ model and BayesCπPED model were small for milk yield and fat content: 0.01 and 0.05, respectively, in the Montbéliarde breed, and 0.02 and 0.01, respectively, in the Holstein breed. The largest increase in the slope for the BayesCπPED model was observed for conception rate: +0.32 in the Montbéliarde breed [i.e., a regression slope closer to 1 for BayesCπ model ($b = 1.31$)] and +0.10 in the Holstein breed, leading to a regression slope closer to 1 for the BayesCπPED model ($b = 0.82$). The best values of slopes were obtained for fat content, with values close to 0.9 for both breeds. No evidence existed of the superiority of the BayesCπPED model over the BayesCπ model either in terms of accuracy or for the regression slope: the simpler model (BayesCπ model) was retained hereafter.

Estimation of Variance Components with the MCMC Algorithm

Figures 2 and 3 display the posterior density of genetic variance (at the top), residual variance (in the middle), and π (at the bottom) during their estimation with the MCMC algorithm in the Holstein (Figure 2) and Montbéliarde (Figure 3) breeds. The sampling is represented according to 200,000 MCMC iterations, with 1 record every 50 cycles and without the burn-in period of 20,000 iterations.

Visual inspection of Figure 2 and the trace plot of the statistical distribution of parameters (genetic and residual variances and π) during their estimation with the MCMC algorithm (results not shown) indicated that the convergence of both genetic and residual variances was almost reached for the 3 studied traits. The posterior distributions covered narrow intervals that were shorter than those defined by the prior distribution. The mean (μ) and standard deviation (σ) for the genetic variance were $\mu_{MY} \approx 396{,}000$ and $\sigma_{MY} \approx 26{,}000$ for milk yield (MY), $\mu_{FC} \approx 8$ and $\sigma_{FC} \approx 0.65$ for fat content (FC), and $\mu_{CR} \approx 45$ and $\sigma_{CR} \approx 3.2$ for con-

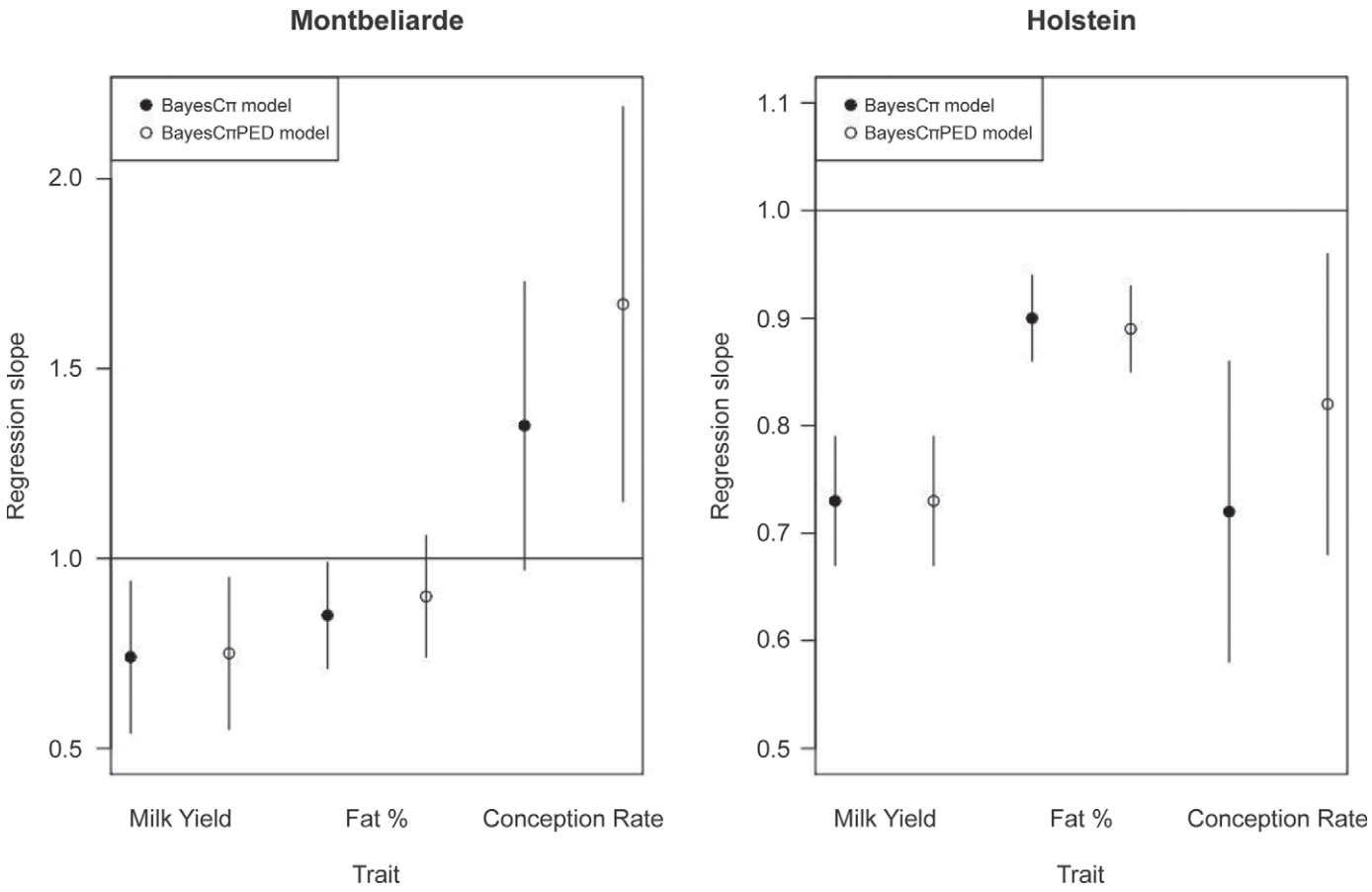

Figure 1. Regression slopes of observed daughter yield deviations (DYD) on predicted DYD in the validation data set without (BayesCπ model) or with (BayesCπPED model) adding a polygenic component to the BayesCπ model.

ception rate (CR). The residual variance gave $\mu_{MY} \approx$ 3,300,000, $\sigma_{MY} \approx$ 180,000, $\mu_{FC} \approx 28$, $\sigma_{FC} \approx 1.6$, $\mu_{CR} \approx$ 2,600, and $\sigma_{CR} \approx 116$. The π parameter was quite accurately estimated and was very low for milk yield and fat content (mean of 0.04 and 0.02 and standard deviation of 0.03 and 0.02, respectively). However, for the conception rate, the convergence of π was not reached, with a mean around 0.5 and a standard deviation of 0.3. Nevertheless, the estimation of both the genetic and residual variances was not affected by the poor estimation of π.

Figure 3 relates to the Montbéliarde breed. The statistical distributions of parameters (results not shown) indicated that the Markov chain was stabilized and appeared constant only for the genetic and residuals variances for the 3 traits. They were very chaotic for the π parameter on milk yield and conception rate. For the genetic variance, $\mu_{MY} \approx 385{,}000$ and $\sigma_{MY} \approx 36{,}000$ for milk yield, $\mu_{FC} \approx 5.5$ and $\sigma_{FC} \approx 0.7$ for fat content, and $\mu_{CR} \approx 31$ and $\sigma_{CR} \approx 4$ for conception rate. The residual variance gave $\mu_{MY} \approx 1{,}600{,}000$, $\sigma_{MY} \approx 396{,}000$, $\mu_{FC} \approx 10$, $\sigma_{FC} \approx 2.6$, $\mu_{CR} \approx 1{,}900$, and $\sigma_{CR} \approx 165$. The value of π stabilized only for fat content, with a value lower than 3%.

Markov chain Monte Carlo chains were also run for the Montbéliarde breed with 1,000,000 iterations and a burn-in of 50,000 iterations (results not shown). The evolution of the estimation of the different variances and π were similar to those observed with 200,000 chains (Figure 3). The correlations between observed and predicted DYD from the validation data set were almost the same, with a maximum difference of ± 0.01 with the model with 200,000 iterations (results not shown). The regression slopes acquired from 1,000,000 iterations were also very close to those obtained with 200,000 iterations. However, the estimation of π fluctuated wildly between chains.

The parameter π was, therefore, arbitrarily set to 10% for the 3 traits in the Montbéliarde breed. Figure 4 shows the posterior density of MCMC chains for genetic variance and residual variance in the Montbéliarde breed for milk yield, fat content, and conception rate, with π fixed at 10%. The trace plot of the statistical distribution of genetic and residual variances during

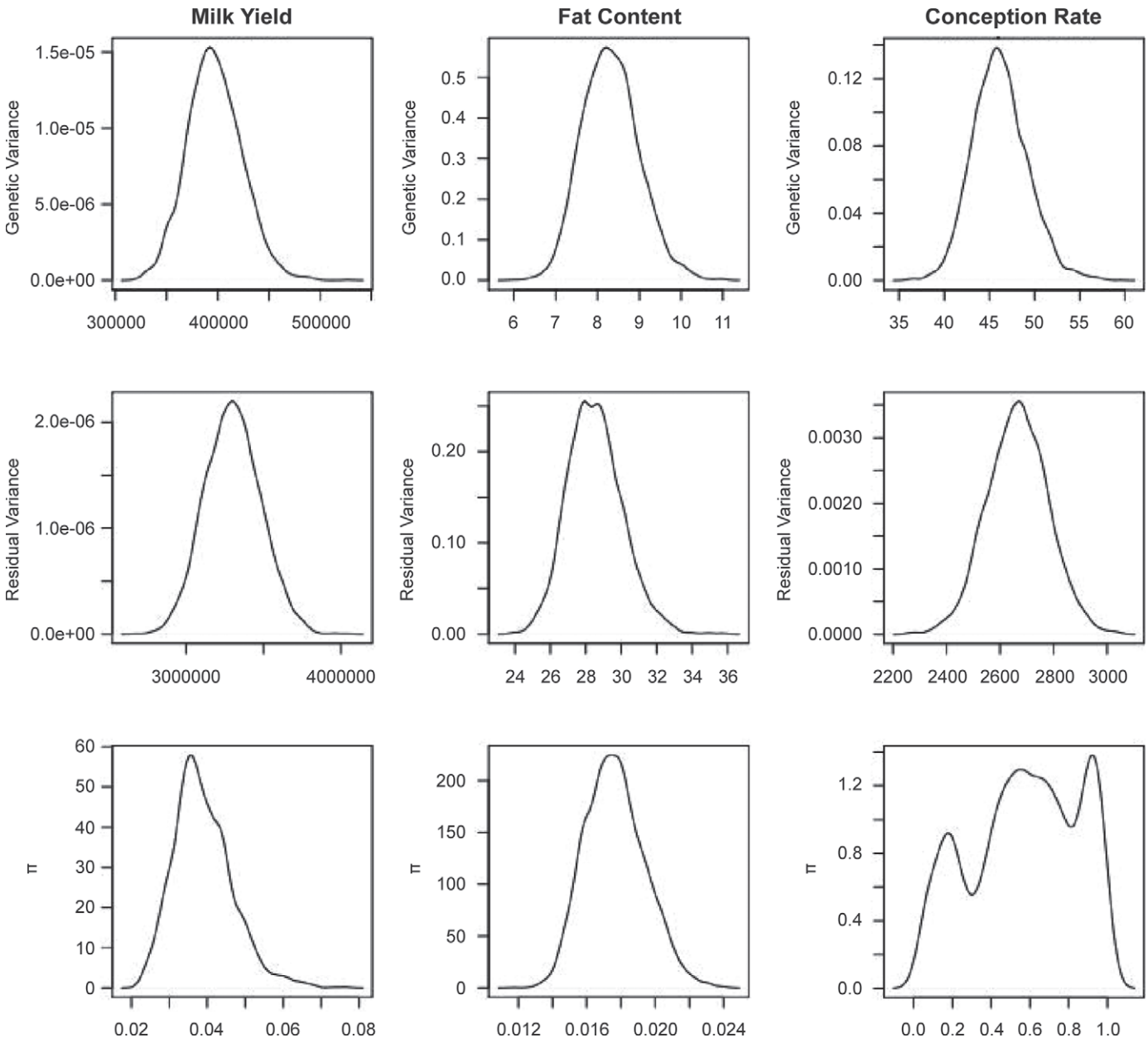

Figure 2. Density of genetic variance, residual variance, and probability π during the Markov chain Monte Carlo (MCMC) algorithm for milk yield, fat content, and conception rate in the Holstein breed.

their estimation with the MCMC algorithm (results not shown) covered narrow intervals. The convergence of both genetic and residual variances was acceptable for the 3 traits. The correlations (ρ) between observed and predicted DYD were exactly the same (ρ = 0.44 for milk yield and ρ = 0.42 for conception rate). On the contrary, for fat content, the correlation decreased from 0.63 to 0.58 but the estimation of genetic variance was more stable. Thereafter, only the BayesCπ model was used without restricting the estimation of π.

Comparison of BayesCπ with Other Methods

Predictive Ability of the Different Methods. Tables 2 and 3 present the correlations (ρ) between observed and predicted DYD from the validation data, for the different methods, in Holsteins and in Montbéliardes, respectively. All correlations were compared one to another using the Hotelling-Williams test with a threshold of 5%. All the methods that rely on genomic information performed significantly better

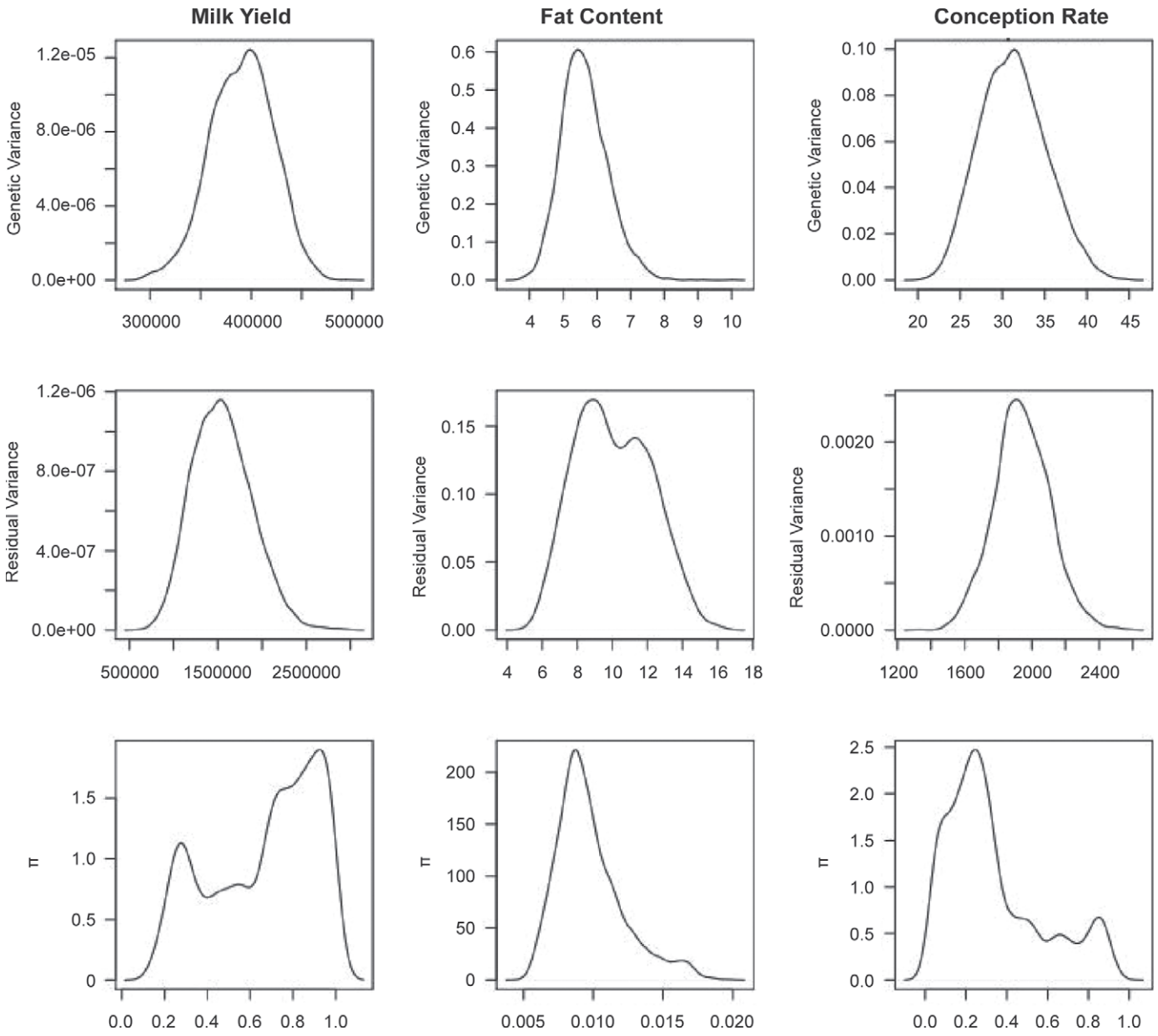

Figure 3. Density of genetic variance, residual variance, and probability π during the Markov chain Monte Carlo (MCMC) algorithm for milk yield, fat content, and conception rate in the Montbéliarde breed.

than pedigree-based BLUP, except with regard to the conception rate (a trait with low heritability) in the Montbéliarde breed, for which no significant difference was observed between any of the correlations. It should be noted that such a high correlation for the pedigree-based BLUP method is not consistent with the lower accuracy of classical BLUP evaluations for conception rate, compared with correlations obtained on the other traits (with higher heritability). This may reflect a high heterogeneity at the genetic level for conception rate among sire families. The correlation of sPLS regression was not significantly different from that of BLUP for the conception rate in Holsteins. In the Montbéliarde breed, the correlations given by all genomic selection methods were not significantly different, except with regard to the fat content, for which GBLUP was significantly less accurate than BayesCπ. Bayesian methods gave the highest correlations, although the difference was nonsignificant, except for fat content in Holsteins. BayesCπ showed a nonsignificant advantage over Bayesian LASSO (+0.01 for milk yield and fat content in Holsteins and +0.09 for fat content in Montbéliardes).

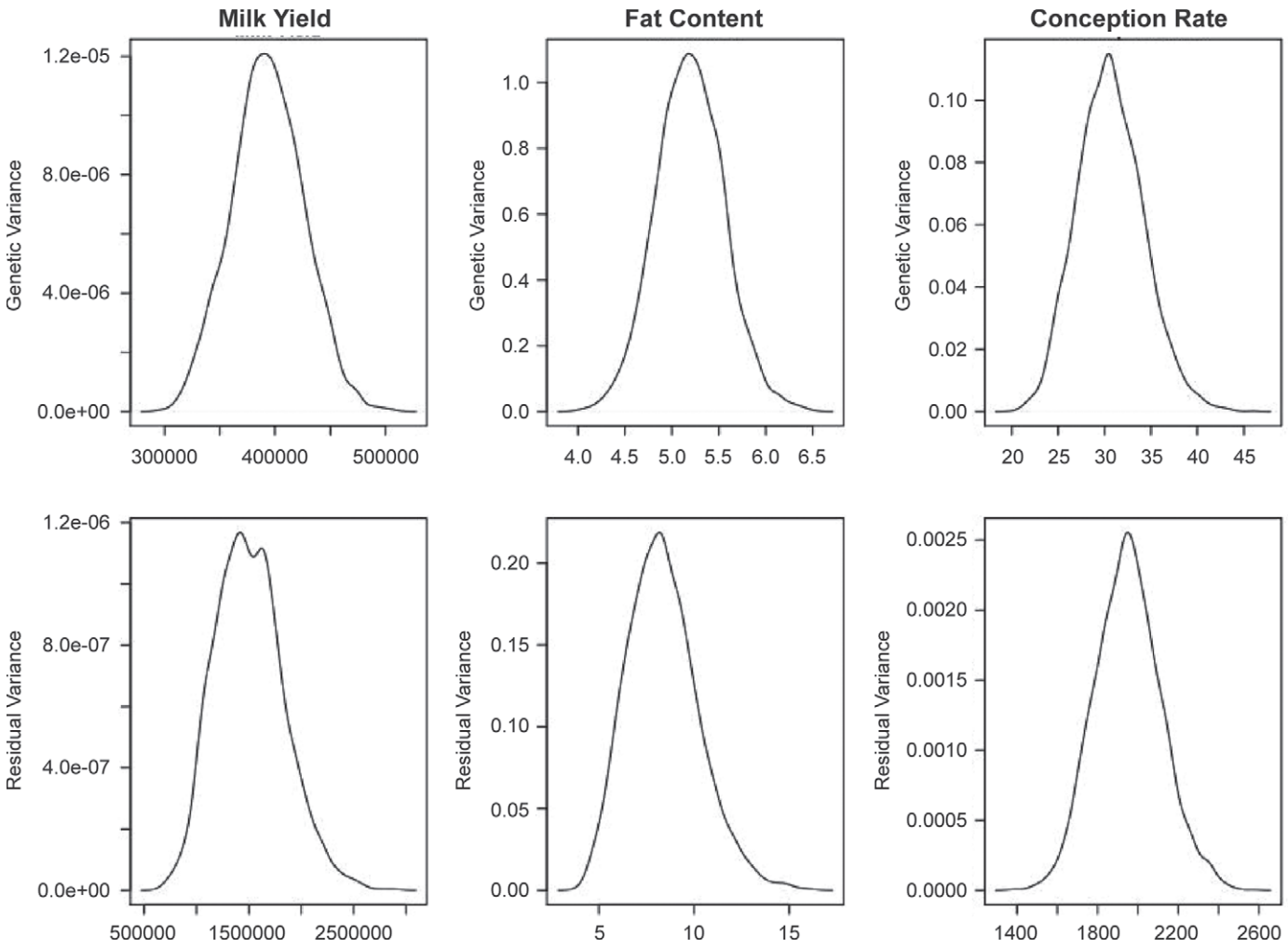

Figure 4. Density of genetic variance and residual variance during the Markov chain Monte Carlo (MCMC) algorithm for milk yield, fat content, and conception rate in the Montbéliarde breed, with $\pi = 10\%$.

The better performances of the Bayesian methods for fat content in Holstein and Montbéliarde breeds is probably due to the effect of the *DGAT1* gene.

Tables 4 and 5 display regression slopes for each of the 3 traits, in Holsteins and Montbéliardes, respectively. A value close to 1 is expected. In Holsteins (Table 4), standard errors were similar for all methods and equal to 0.03, 0.02, and 0.07 for milk yield, fat content, and conception rate, respectively. Genomic selection methods provided lower regression slopes than pedigree-based BLUP. Among the genomic selection methods for predicting milk yield and fat content,

Table 2. Correlations between observed daughter yield deviations (DYD) and predicted DYD in the validation data set provided by pedigree-based BLUP, genomic BLUP (GBLUP), partial least squares (PLS) regression, sparse PLS (sPLS) regression, Bayesian least absolute shrinkage and selection operator (LASSO), and BayesCπ models in the Holstein breed

	Model					
Item	BLUP	GBLUP	PLS	sPLS	Bayesian LASSO	BayesCπ
Milk yield	0.38	0.56	0.53	0.48	0.56	0.57
Fat %	0.44	0.72	0.70	0.66	0.79	0.80
Conception rate	0.28	0.35	0.33	0.29	0.34	0.34

Table 3. Correlations between observed daughter yield deviations (DYD) and predicted DYD in the validation data set provided by pedigree-based BLUP, genomic BLUP (GBLUP), partial least squares (PLS) regression, sparse PLS (sPLS) regression, Bayesian least absolute shrinkage and selection operator (LASSO), and BayesCπ models in the Montbéliarde breed

	Model					
Item	BLUP	GBLUP	PLS	sPLS	Bayesian LASSO	BayesCπ
Milk yield	0.28	0.42	0.44	0.38	0.44	0.44
Fat %	0.40	0.52	0.58	0.56	0.53	0.62
Conception rate	0.43	0.47	0.43	0.43	0.43	0.43

the Bayesian methods were the most efficient, with a small advantage for Bayesian LASSO (+0.01 for milk yield and +0.03 for fat content). The PLS regression methods were the least efficient. With regard to the conception rate, GBLUP gave a slope value close to that obtained with BLUP: 0.78 with GBLUP and 0.80 with BLUP. In Montbéliardes (Table 5), standard errors were similar for all methods and equal to 0.10, 0.07, and 0.20 for milk yield, fat content, and conception rate, respectively. The PLS regression methods were also shown to be the least-efficient methods, except for the prediction of fat content, for which the regression slope for PLS regression equaled the regression slopes for BLUP and GBLUP (±0.02 away from 1). The slope obtained with GBLUP for the milk yield was 0.84; that is, +0.10 compared with the regression slopes with Bayesian methods and BLUP. Finally, very bad results were obtained for the conception rate in Montbéliardes (1.35 with BayesCπ compared with 2.27 with sPLS regression), but standard errors were very large.

Estimation of SNP Effects. Figures 5 and 6 show the estimation of SNP effects (in genetic standard deviation units) for Bayesian LASSO and BayesCπ, and *VIP* coefficients for PLS and sPLS regressions. Emphasis was placed on the position of the SNP with the largest effects so that *VIP* coefficients and genetic standard deviation units could be compared. All SNP were represented on the graphs, even SNP with zero effect. The PLS and sPLS regressions gave almost the same positions for important SNP, so only the results obtained with sPLS regression are shown. Moreover, the variable selection performed by sPLS regression allows a simpler interpretation of *VIP* coefficients (Colombani et al., 2010).

When BayesCπ with Bayesian LASSO were compared, the positions of the most important SNP were found to be similar for most cases. For milk yield in Holsteins, a genome region on chromosome 5 was particularly highlighted with BayesCπ, but represented a very small peak in Bayesian LASSO. However, chromosome 14 stood out strongly with both Bayesian LASSO and BayesCπ and almost the same SNP were selected. For fat content, Bayesian LASSO weighted 1 SNP particularly on chromosome 14 that resulted in a weaker effect for the other SNP. This 1 SNP on chromosome 14 was also the most weighted with BayesCπ and sPLS regression, but with a lesser difference between the effect of the first and the second most important SNP. For conception rate, the graphs were similar with BayesCπ and Bayesian LASSO, with a large number of peaks. Sparse PLS regression provided similar results to BayesCπ: the most important peaks of BayesCπ being also strongly highlighted in sPLS regression. The SNP with the largest effects were almost the same with sPLS regression and BayesCπ, except for 1 SNP on chromosome 21 that showed up strongly for milk yield, but only with BayesCπ.

In Montbéliardes, almost identical graphs were observed with BayesCπ and Bayesian LASSO with regard to the position of the peaks and the size of the effects

Table 4. Regression slopes of observed daughter yield deviations (DYD) on predicted DYD in the validation data set provided by pedigree-based BLUP, genomic BLUP (GBLUP), partial least squares (PLS) regression, sparse PLS (sPLS) regression, Bayesian least absolute shrinkage and selection operator (LASSO), and BayesCπ models in the Holstein breed

	Model					
Item	BLUP	GBLUP	PLS	sPLS	Bayesian LASSO	BayesCπ
Milk yield	0.79	0.68	0.65	0.53	0.74	0.73
Fat %	0.97	0.87	0.80	0.69	0.93	0.90
Conception rate	0.80	0.78	0.60	0.54	0.72	0.72

Table 5. Regression slopes of observed daughter yield deviations (DYD) on predicted DYD in the validation data set provided by pedigree-based BLUP, genomic BLUP (GBLUP), partial least squares (PLS) regression, sparse PLS (sPLS) regression, Bayesian least absolute shrinkage and selection operator (LASSO), and BayesCπ models in the Montbéliarde breed

	Model					
Item	BLUP	GBLUP	PLS	sPLS	Bayesian LASSO	BayesCπ
Milk yield	0.74	0.84	0.64	0.63	0.74	0.74
Fat %	1.01	1.01	0.98	0.81	0.91	0.85
Conception rate	1.78	1.76	1.79	2.27	1.36	1.35

for milk yield and conception rate. However, for fat content, Bayesian LASSO identified only chromosome 14, but with a much smaller estimated effect than with BayesCπ, indicating that the shrinkage of SNP effect was greater with Bayesian LASSO than BayesCπ. All peaks detected with BayesCπ were found with sPLS regression at the same position but with a different ranking for all traits. Moreover, for milk yield, some regions (such as chromosome 7 and 15) were detected as having a strong effect with sPLS regression but not with BayesCπ. For fat content, the same peaks appeared with both BayesCπ and sPLS regression, but sPLS regression included more SNP in its peaks than BayesCπ. For the conception rate, the effect peaks were found at the same positions with all methods, but were more accentuated in sPLS regression than in Bayesian methods.

DISCUSSION

The objective of this study was to analyze the predictive ability of Bayesian LASSO and BayesCπ methods, in particular, in comparison with other methods used in the genomic evaluation of dairy cattle. Our first step was to explore the BayesCπ method with different settings of the model, considering the inclusion of pedigree information and the handling of the π value. The results of this step of the study using real data (Holstein and Montbéliarde breeds) demonstrated that the addition of a polygenic component to the BayesCπ model or setting the value of π did not, in most cases, improve the correlation between observed phenotypes and GEBV. Therefore, we retained the simplest BayesCπ model. The predictive ability of BayesCπ was then compared with another Bayesian method (Bayesian LASSO), BLUP approaches (pedigree-based BLUP and GBLUP), and dimension reduction methods (PLS regression and its variable selection variant, sPLS regression). Pedigree-based BLUP was less accurate than genomic selection methods but provided better regression slopes. However, the different ways of comparing genomic selection methods failed to demonstrate the systematic superiority of BayesCπ or any other approach.

Polygenic Effects and Genomic Selection

The results of the present study show that genomic selection methods are more accurate than pedigree-based BLUP, but with a limited gain of accuracy. If linkage disequilibria exist between SNP and QTL and sufficient records exist in the reference set to estimate SNP effects accurately, GEBV accuracy is higher than pedigree-based EBV (Meuwissen et al., 2001; Habier et al., 2007). This is due to the fact that the accuracies of GEBV estimated using a genomic model without pedigree information are affected by the genetic relationship among individuals of the reference population. Habier et al. (2007) demonstrated that SNP markers are able to capture genetic relationships among genotyped animals. Habier et al. (2010) also tested the effect on the accuracy of GEBV of different values of maximum additive genetic relationship (a_{max}) between bulls in training and validation populations. They showed that the accuracy of GEBV were the highest with $a_{max} = 0.6$ (i.e., a strong relationship between training and validation bulls), but the gain of genomic selection over pedigree-based BLUP was less than with smaller values of a_{max}. When the relationships between training and validation bulls were high, pedigree-based BLUP performed better and so the gain of genomic selection was smaller. The results of our study are in agreement with the conclusion of Habier et al. (2010). Our training data sets contained sires, full sibs, and half sibs of the bulls in the validation set and the relationships between the bulls of our reference population were high, so pedigree-based BLUP was quite efficient and the gain of genomic selection methods over pedigree-based BLUP was limited (the mean gain of correlations of GBLUP over BLUP was equal to 0.18 in Holsteins and 0.10 in Montbéliardes).

The correlations between observed DYD and predicted DYD in the validation data set were very close with

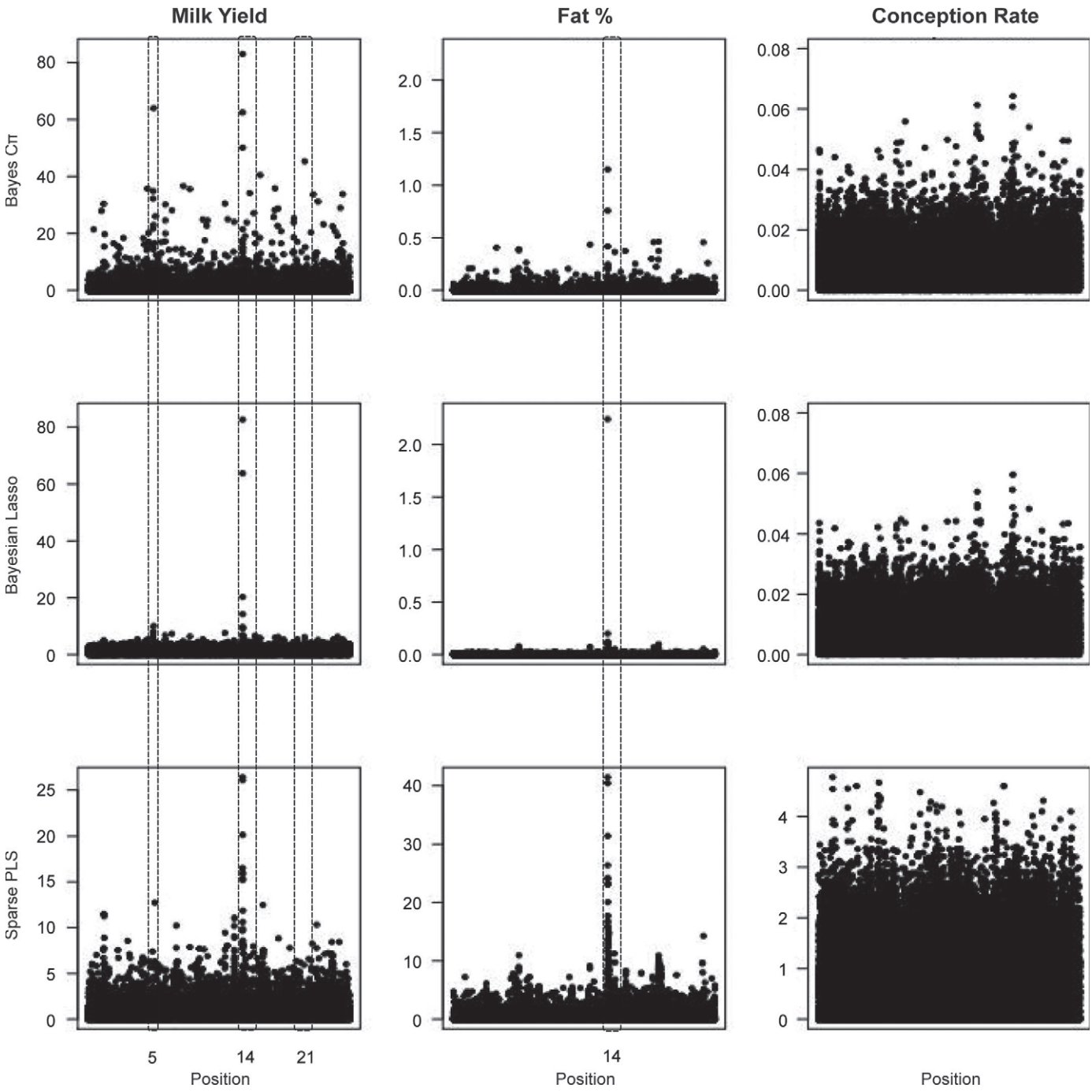

Figure 5. Estimation of SNP effects by BayesCπ (BayesCπ model), Bayesian least absolute shrinkage and selection operator (LASSO), and variable importance in projection (*VIP*) coefficients for sparse partial least squares (sPLS) regression for milk yield, fat content, and conception rate in the Holstein breed according to the marker position along the genome.

the BayesCπ model (including only SNP information) and BayesCπPED model (including both polygenic and SNP effects), whatever the trait or the breed (Table 1). The regression slopes were also very similar for both models (Figure 1). Mrode et al. (2011) presented a study that aimed at testing the inclusion of polygenic effects using 11,480 Holstein-Friesian bulls in the United Kingdom in a linear model equivalent to a GBLUP model. They showed that the correlations for production traits decreased slightly but the regression coefficient increased by approximately 0.1 for all traits. Liu et al. (2011) showed that including a polygenic effect

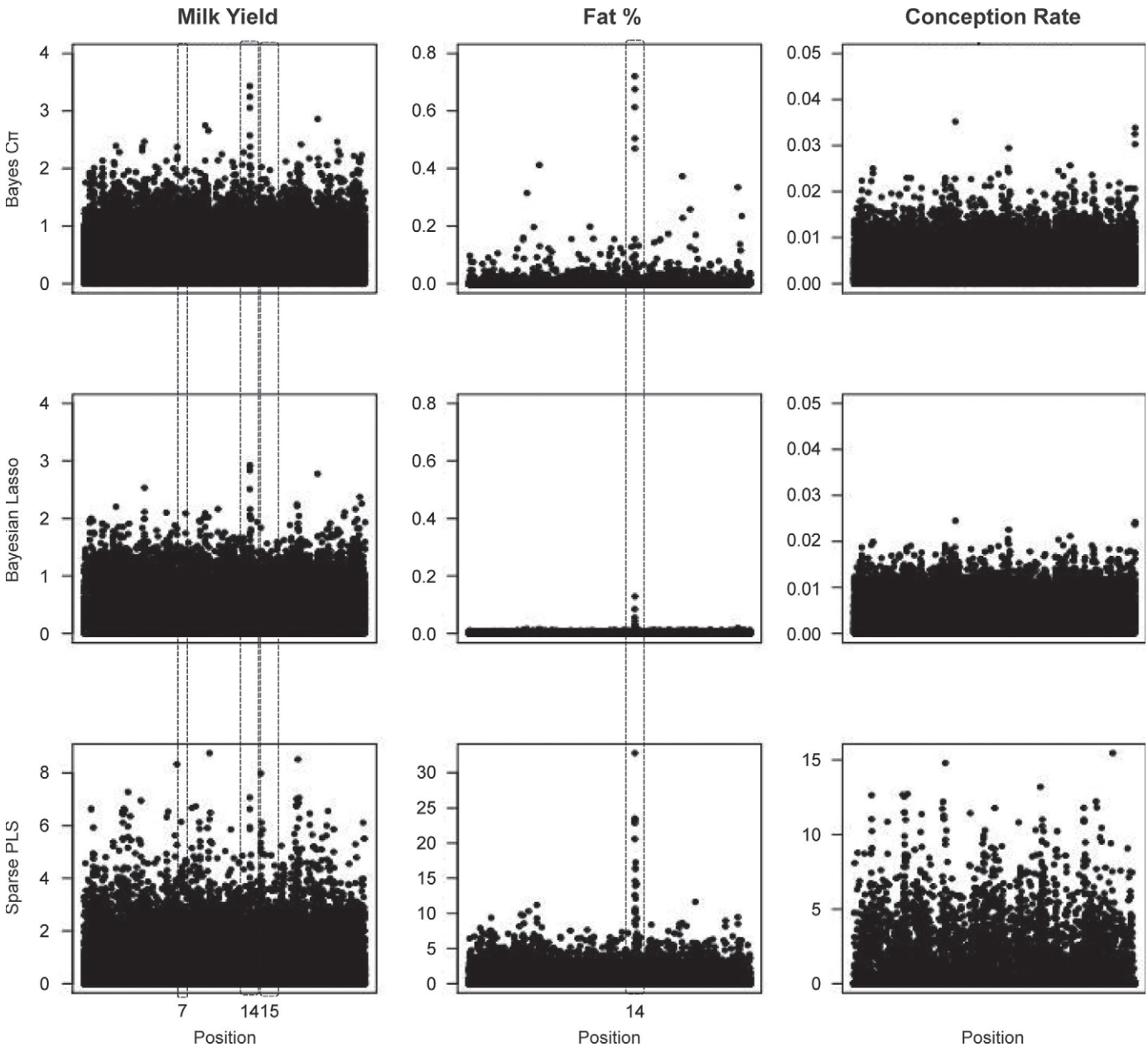

Figure 6. Estimation of SNP effects by BayesCπ (BayesCπ model), Bayesian least absolute shrinkage and selection operator (LASSO), and variable importance in projection (*VIP*) coefficients in sparse partial least squares (sPLS) for milk yield, fat content, and conception rate in the Montbéliarde breed according to the marker position along the genome.

in a GBLUP model resulted in decreased correlations between direct genomic values and EBV. In our study, the inclusion of polygenic effects also led to slightly smaller correlations for fat content in both breeds but with no significant differences. These results show that SNP marker information could contain a part of pedigree information. In Lacaune dairy sheep, similar conclusions were obtained with a reference population of approximately 2,500 proven rams and 44,000 SNP (Robert-Granié et al., 2011). Inclusion of infinitesimal effects in the prediction model with BayesCπ had little effect on accuracies and led to slightly better slopes of regressions.

However, a difference between the BayesCπ model and BayesCπPED model appeared regarding the number of SNP selected for all traits in Montbéliardes and for the conception rate in Holsteins. The difference between the number of SNP included in the BayesCπ model and

BayesCπPED model was about −5,000 SNP for milk yield and +9,000 SNP for conception rate in Montbéliardes and +6,000 SNP for conception rate in Holsteins. However, the graphs of the posterior distribution of π during the MCMC algorithm (Figures 2 and 3) showed that it was difficult to obtain a good convergence of π in these cases. So, it was problematic to obtain an accurate estimation of the parameter π in these cases. The final value of π varied around the mean (i.e., 0.5) and the number of SNP selected fluctuated between 15,000 and 25,000 in these cases. For milk yield and fat content in Holsteins, the number of SNP selected was stable between the 2 models: approximately 1,600 SNP for milk yield and 800 SNP for fat content. In the Montbéliarde breed, the result was more surprising for fat content because π seemed to be well estimated during the MCMC process (around 0.01 in the BayesCπ model but around 0.51 in the BayesCπPED model; i.e., a difference of +19,000 SNP between these 2 models). The inclusion of polygenic effects (BayesCπPED model) for fat content in Montbéliardes interfered with SNP selection in BayesCπ: the model could not identify the most important SNP for the prediction of this trait, although no change occurred in the correlation between observed DYD and predicted DYD in the validation set. Several situations were tested in which the weight allocated to the polygenic effects was modified, but the correlation between observed and predicted DYD was the same.

Comparison of Methods

The results obtained with MCMC chains in the BayesCπ model were compared with Bayesian LASSO, GBLUP, and PLS and sPLS regressions. Hayes et al. (2009) and VanRaden et al. (2009) presented 2 reviews of empirical results in dairy cattle that pointed out the similarity of GBLUP and BayesB, as far as predictive ability is concerned. Croiseau et al. (2011) compared the elastic net approach to GBLUP on the French data sets used in this study. They obtained the same correlations as BayesCπ with better regression slopes. Our results are in good agreement with those studies, as the Bayesian methods reached the same accuracies as GBLUP for most traits. The regression slopes did not allow differentiating between GBLUP and Bayesian methods either. The PLS regression variants were the least efficient, both in regard to correlation and regression coefficients in most cases.

The regression slopes of observed DYD on estimated DYD were less than 1 in most cases. This is probably due to the fact that the reference populations were made up of strongly selected bulls. Vitezica et al. (2011) proposed that the strong selection of animals in dairy cattle schemes and, therefore, on the animals of the reference populations, may result in the observed biases of the regression coefficients. Biases might also be introduced by the use of DYD, as suggested by Patry and Ducrocq (2011).

BayesCπ and Bayesian LASSO seem to follow the same pattern in the present study. The difference between GBLUP and BayesCπ was high for fat content both in Montbéliarde and Holstein breeds. To obtain a good idea of the genome regions involved in the prediction equation of each method, we studied the graphs of SNP effects. The genome areas with the largest effects were almost identical with BayesCπ, Bayesian LASSO and sPLS regression, whatever the trait or breed. The graphs for conception rate were similar for the Bayesian methods but showed some dissimilarity with sPLS regression: the peaks were found at the same positions but the ranking of these peaks was different. However, this could be explained by the fact that the conception rate seems to be a very polygenic trait and that its low heritability leads to limited prediction accuracy with all methods. One can note that the number of SNP with nonzero effects is relatively small for fat content and this in both breeds. A specific peak stood out particularly with all methods for the fat content at the beginning of chromosome 14. In Holsteins, this genome region corresponds to the *DGAT1* gene (Grisart et al., 2004), which, when mutated, has a major effect on the fat content in milk. Hence, the superior accuracy of BayesCπ and Bayesian LASSO against GBLUP ($\rho_{BAYESC\pi} = 0.80$, $\rho_{BayesianLASSO} = 0.79$, but $\rho_{GBLUP} = 0.72$) for the fat content trait in Holsteins could be explained by the small number of QTL related to this trait. For the fat content trait in Montbéliardes, BayesCπ outperformed GBLUP ($\rho_{BAYESC\pi} = 0.62$ and $\rho_{GBLUP} = 0.52$), but surprisingly, Bayesian LASSO was closer to GBLUP than BayesCπ, with $\rho_{BayesianLASSO} = 0.53$. Concerning SNP effects, Bayesian LASSO highlighted only one region on chromosome 14, whereas BayesCπ and sPLS regression highlighted 4 or 5 regions. Therefore, the equation of prediction in Bayesian LASSO was based on a very small number of SNP, all positioned within the same genome region, and this seemed to affect the accuracy of prediction. For milk yield in Holsteins, a few SNP were retained by BayesCπ (about 1,500 SNP) but almost 23,000 SNP were necessary with sPLS regression. This suggests that milk yield is affected by a large number of QTL. The results of GBLUP and Bayesian methods for milk yield in Holsteins were almost identical.

The conclusions established in this work should be transferable to other studies if the characteristics of the traits studied are considered correctly. Bayesian

methods seem to perform well, whatever the trait or population, but in some cases, GBLUP is as accurate as Bayesian methods. These results were confirmed on the study of 3 traits in Lacaune sheep breed (Duchemin et al., 2012), indicating the superiority of BayesCπ. Regarding computing time, Bayesian methods are the less efficient, with about 12 h per trait in our rather small data sets. Genomic BLUP requires the inversion of the genomic relationship matrix for all traits, which took about 1 h for our data set. Then, once the genomic relationship matrix was inverted, computation was a matter of seconds.

CONCLUSIONS

The first goal of this study was to explore the predictive ability of BayesCπ in a genomic evaluation context. According to the accuracy and regression slope, the inclusion of pedigree information in the BayesCπ model did not change the results, nor did fixing the value of π. BayesCπ did not show a large advantage over other methods, except for traits with a small final number of selected SNP such as fat content. No genomic selection method tested in this study outperformed the others, but it is interesting to note that the position of the SNP selected by the different models (LASSO, BayesCπ, and sPLS) were close. The next step of our work will be to compare more precisely the SNP selected by all the methods and to study whether the genome regions highlighted by some genomic selection methods correspond to the QTL detected by specific QTL detection methods (i.e., linkage analysis, linkage disequilibrium, or linkage disequilibrium-linkage analysis).

ACKNOWLEDGMENTS

This work was supported by the French project AMASGEN (2009–2011), financed by the French National Research Agency (ANR, Paris, France) and Apis Gène (Paris, France). Labogena (Jouy-en-Josas, France; http://www.labogena.fr/) is gratefully acknowledged for providing the genotypes. The project was partly supported by the Toulouse Midi-Pyrénées bioinformatics platform GenoToul (Toulouse, France; http://bioinfo.genotoul.fr/).

REFERENCES

Boichard, D., and E. Manfredi. 1994. Genetic analysis of conception rate in French Holstein cattle. Acta Agric. Scand. A Anim. Sci. 44:138–145.

Chun, H., and S. Keleş. 2009. Expression quantitative trait loci mapping with multivariate sparse partial least squares regression. Genetics 182:79–90.

Cole, J. B., P. M. VanRaden, J. R. O'Connell, C. P. Van Tassell, T. S. Sonstegard, R. D. Schnabel, J. F. Taylor, and G. R. Wiggans. 2009. Distribution and location of genetic effects for dairy traits. J. Dairy Sci. 92:2931–2946.

Colombani, C., P. Croiseau, S. Fritz, F. Guillaume, A. Legarra, V. Ducrocq, and C. Robert-Granié. 2012. A comparison of partial least squares (PLS) and sparse PLS regressions in genomic selection in French dairy cattle. J. Dairy Sci. 95:2120–2131.

Colombani, C., A. Legarra, P. Croiseau, F. Guillaume, S. Fritz, V. Ducrocq, and C. Robert-Granié. 2010. Application of PLS and sparse PLS regression in genomic selection. In 9th World Congr. Genet. Appl. Livest. Prod., Leipzig, Germany. Gesellschaft für Tierzuchtwissenschaften eV, Gießen, Germany.

Coster, A., J. W. M. Bastiaansen, M. P. L. Calus, J. A. M. van Arendonk, and H. Bovenhuis. 2010. Sensitivity of methods for estimating breeding values using genetic markers to the number of QTL and distribution of QTL variance. Genet. Sel. Evol. 42:9.

Croiseau, P., A. Legarra, F. Guillaume, S. Fritz, A. Baur, C. Colombani, C. Robert-Granié, D. Boichard, and V. Ducrocq. 2011. Fine tuning genomic evaluations in dairy cattle through SNP preselection with the elastic-net algorithm. Genet. Res. (Camb.) 93:409–417.

de los Campos, G., H. Naya, D. Gianola, J. Crossa, A. Legarra, E. Manfredi, K. Weigel, and J. M. Cotes. 2009. Predicting quantitative traits with regression models for dense molecular markers and pedigree. Genetics 182:375–385.

Druet, T., and M. Georges. 2010. A hidden Markov model combining linkage and linkage disequilibrium information for haplotype reconstruction and quantitative trait locus fine mapping. Genetics 184:789–798.

Duchemin, S. I., C. Colombani, A. Legarra, G. Baloche, H. Larroque, J.-M. Astruc, F. Barillet, C. Robert-Granié, and E. Manfredi. 2012. Genomic selection in the French Lacaune dairy sheep breed. J. Dairy Sci. 95:2723–2733.

Gianola, D., G. de los Campos, W. G. Hill, E. Manfredi, and R. Fernando. 2009. Additive genetic variability and the Bayesian alphabet. Genetics 183:347–363.

Gredler, B., K. G. Nirea, T. R. Solberg, C. Egger-Danner, T. H. E. Meuwissen, and J. Sölkner. 2009. Genomic selection in Fleckvieh/Simmental—First results. In Proc. Interbull Mtg., Barcelona, Spain. Interbull Centre, Dept. Animal Breeding and Genetics, Swedish University of Agriculture Sciences (SLU), Uppsala, Sweden.

Grisart, B., F. Farnir, L. Karim, N. Cambisano, J.-J. Kim, A. Kvasz, M. Mni, P. Simon, J.-M. Frère, W. Coppieters, and M. Georges. 2004. Genetic and functional confirmation of the causality of the *DGAT1 K232A* quantitative trait nucleotide in affecting milk yield and composition. Proc. Natl. Acad. Sci. USA 101:2398–2403.

Habier, D., R. L. Fernando, and J. C. M. Dekkers. 2007. The impact of genetic relationship information on genome-assisted breeding values. Genetics 177:2389–2397.

Habier, D., R. L. Fernando, K. Kizilkaya, and D. J. Garrick. 2011. Extension of the Bayesian alphabet for genomic selection. BMC Bioinformatics 12:186.

Habier, D., J. Tetens, F.-R. Seefried, P. Lichtner, and G. Thaller. 2010. The impact of genetic relationship information on genomic breeding values in German Holstein cattle. Genet. Sel. Evol. 42:5.

Harris, B. L., D. L. Johnson, and R. J. Spelman. 2009. Genomic selection in New Zealand and the implications for national genetic evaluation. Pages 325–330 in Identification, Breeding, Production, Health and Recording of Farm Animals. Proceedings of the 36th ICAR Biennial Session, Niagara Falls. International Committee for Animal Recording (ICAR), Rome, Italy.

Hayes, B. J. 2009. Genomic selection in the era of the $1000 genome sequence. In Proc. Symposium Statistical Genetics of Livestock for the Post-Genomic Era, Madison, WI. Department of Dairy Science, University of Wisconsin-Madison.

Hayes, B. J., A. J. Chamberlain, S. Maceachern, K. Savin, H. McPartlan, I. MacLeod, L. Sethuraman, and M. E. Goddard. 2009. A genome map of divergent artificial selection between *Bos taurus* dairy cattle and *Bos taurus* beef cattle. Anim. Genet. 40:176–184.

Kizilkaya, K., R. L. Fernando, and D. J. Garrick. 2010. Genomic prediction of simulated multibreed and purebred performance using

observed fifty thousand single nucleotide polymorphism genotypes. J. Anim. Sci. 88:544–551.

Lê Cao, K.-A., I. González, and S. Déjean. 2009. integrOmics: An R package to unravel relationships between two omics datasets. Bioinformatics 25:2855–2856.

Lê Cao, K.-A., D. Rossouw, C. Robert-Granié, and P. Besse. 2008. A sparse PLS for variable selection when integrating omics data. Stat. Appl. Genet. Mol. Biol. 7:35.

Legarra, A., C. Robert-Granié, P. Croiseau, F. Guillaume, and S. Fritz. 2011. Improved Lasso for genomic selection. Genet. Res. (Camb.) 93:77–87.

Liu, Z., F. R. Seefried, F. Reinhardt, S. Rensing, G. Thaller, and R. Reents. 2011. Impacts of both reference population size and inclusion of a residual polygenic effect on the accuracy of genomic prediction. Genet. Sel. Evol. 43:19.

Long, N., D. Gianola, J. M. Rosa, and K. A. Weigel. 2011. Dimension reduction and variable selection for genomic selection: Application to predicting milk yield in Holsteins. J. Anim. Breed. Genet. 128:247–257.

Meuwissen, T. H. E., B. J. Hayes, and M. E. Goddard. 2001. Prediction of total genetic value using genome-wide dense marker maps. Genetics 157:1819–1829.

Mevik, B.-H., and H. R. Cederkvist. 2004. Mean squared error of prediction (MSEP) estimates for principal component regression (PCR) and partial least squares regression (PLSR). J. Chemometr. 18:422–429.

Moser, G., B. Tier, R. Crump, M. Khatkar, and H. Raadsma. 2009. A comparison of five methods to predict genomic breeding values of dairy bulls from genome-wide SNP markers. Genet. Sel. Evol. 41:56.

Mrode, R. A., T. Krzyzelewski, K. Moore, M. Winters, and M. Coffey. 2011. The implementation of genomic evaluations in the UK. In Proc. of Interbull meeting, Stavanger, Norway. Interbull Centre, Dept. Animal Breeding and Genetics, Swedish University of Agriculture Sciences (SLU), Uppsala, Sweden.

Mrode, R. A., and G. J. Swanson. 2004. Calculating cow and daughter yield deviations and partitioning of genetic evaluations under a random regression model. Livest. Prod. Sci. 86:253–260.

Ostersen, T., O. F. Christensen, M. Henryon, B. Nielsen, G. Su, and P. Madsen. 2011. Deregressed EBV as the response variable yield more reliable genomic predictions than traditional EBV in purebred pigs. Genet. Sel. Evol. 43:38.

Patry, C., and V. Ducrocq. 2011. Evidence of biases in genetic evaluations due to genomic preselection in dairy cattle. J. Dairy Sci. 94:1011–1020.

Robert-Granié, C., S. Duchemin, H. Larroque, G. Baloche, F. Barillet, C. Moreno, A. Legarra, and E. Manfredi. 2011. A comparison of various methods for the computation of genomic breeding values in French Lacaune dairy sheep. In Proc. EAAP Annual Meeting, Stavanger, Norway. Wageningen Academic Publishers, Wageningen, the Netherlands.

Shen, H. P., and J. H. Z. Huang. 2008. Sparse principal component analysis via regularized low rank matrix approximation. J. Multivariate Anal. 99:1015–1034.

Solberg, T. R., A. K. Sonesson, J. A. Woolliams, and T. H. E. Meuwissen. 2009. Reducing dimensionality for prediction of genome-wide breeding values. Genet. Sel. Evol. 41:29.

Sorensen, D., and D. Gianola. 2002. Likelihood, Bayesian and MCMC Methods in Quantitative Genetics. Springer, New York, NY.

Steiger, J. H. 1980. Tests for comparing elements of a correlation matrix. Psychol. Bull. 87:245–251.

Sun, X., D. Habier, R. L. Fernando, D. J. Garrick, and J. C. M. Dekkers. 2011. Genomic breeding value prediction and QTL mapping of QTLMAS2010 data using Bayesian methods. BMC Proceedings 5(Suppl. 3):S13.

Tibshirani, R. 1996. Regression shrinkage and selection via the Lasso. J. Royal Stat. Soc. B 58:267–288.

VanRaden, P. M. 2008. Efficient methods to compute genomic predictions. J. Dairy Sci. 91:4414–4423.

VanRaden, P. M., C. P. Van Tassell, G. R. Wiggans, T. S. Sonstegard, R. D. Schnabel, J. F. Taylor, and F. S. Schenkel. 2009. Invited review: Reliability of genomic predictions for North American Holstein bulls. J. Dairy Sci. 92:16–24.

VanRaden, P. M., and G. R. Wiggans. 1991. Derivation, calculation, and use of national animal model information. J. Dairy Sci. 74:2737–2746.

Van Sickle, J. 2003. Analyzing correlations between stream and watershed attributes. J. Am. Water Resour. Assoc. 39:717–726.

Vitezica, Z. G., I. Aguilar, I. Misztal, and A. Legarra. 2011. Bias in genomic predictions for populations under selection. Genet. Res. (Camb.) 93:357–366.

Weigel, K. A., G. de los Campos, O. González-Recio, H. Naya, X. L. Wu, N. Long, G. J. M. Rosa, and D. Gianola. 2009. Predictive ability of direct genomic values for lifetime net merit of Holstein sires using selected subsets of single nucleotide polymorphism markers. J. Dairy Sci. 92:5248–5257.

Wold, H. 1966. Estimation of principal components and related models by iterative least squares. Pages 391–420 in Multivariate Analysis. P. R. Krishnaiah, ed. Academic Press, New York, NY.